Homework3
CS157b
Craig Huff

1) For each of the following operations, write an iterator that uses an algorithm described in class to enumerate the output of the following operations:

   (a) Distinct (R)
   Open: Start with R.Open( ). Then, create an empty hash table, *S,* that will represent a set of tuples of *R* seen so far.
   GetNext: Repeat R.GetNext( )until all rows have been scanned or a tuple *t* that is not in set *S* is returned. If *t* is not in S, then insert *t* into *S* and return that tuple.
   Close: Once all tuples have been scanned, R.Close( ).

   (b) Bag Difference (R1, R2)
   Open: Start with R1.Open( ). Create an empty set, S, to store all tuples that are in R.
   GetNext: As long as R1.GetNext( ) returns a tuple *t*, we will add *t* to S. Once the table has no more values, we call R1.Close( ) and then R2.Open( ). Then, while R2.GetNext( ) returns a tuple *t,* we check for the first instance of it in S. If it exists in S, remove the first instance of it. If it doesn't exist, don't remove anything.
   Close:
   Once all tuples in R2 have been scanned, R2.Close( ).

2) If *B*(*R*)=*B*(*S*)=25,000 and *M*=2000, what are the disk I/O requirements of:

   (a) two-pass set intersection from class

   3(B(R) + B(S)) Disk I/O's **The second pass will only work if B(S)+B(R) ≤ M^2
   3(25,000 + 25,000) = 150,000 Disk I/O's **AND** 50,000 ≤ 4,000,000 ✓

   (b) sort-join from class.

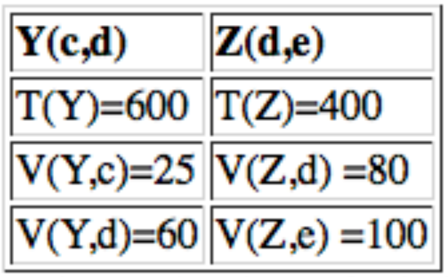
   Assuming this is a 2-two pass simple sort join this will take…
   5(B(R) + B(S)) Disk I/O's **The second pass will only work if B(S)+B(R) ≤ M^2
   5(25,000 + 25,000) = 250,000 Disk I/O's **AND** 50,000 ≤ 4,000,000 ✓

3) Come up with additional query parsing rules to add to our rules to handle SQL join clauses. (I'm assuming these are in a SWF query)

   Inner Join Rule: <SWF> ::= SELECT <SelList> FROM <FromList> INNER JOIN <Relation> ON <Condition> WHERE <Condition>
   Left Outer Join Rule: <SWF> ::= SELECT <SelList> FROM <FromList> LEFT JOIN <Relation> ON <Condition> WHERE <Condition>

Right Outer Join Rule: <SWF> ::= SELECT <SelList> FROM <FromList> RIGHT JOIN <Relation> ON <Condition> WHERE <Condition>
Full Outer Join Rule: <SWF> ::= SELECT <SelList> FROM <FromList> FULL JOIN <Relation> ON <Condition> WHERE <Condition>
Self Join Rule: <SWF> ::= SELECT <SelList> FROM <FromList>, <FromList> WHERE <Condition>

4) Estimate the sizes of relations that are the results from the following queries:



Y(c,d)	Z(d,e)
T(Y)=600	T(Z)=400
V(Y,c)=25	V(Z,d) =80
V(Y,d)=60	V(Z,e) =100

(a)$\sigma_{c=50}(Y)$

$$\frac{T(Y)}{V(Y,c)} = \frac{600}{25} = 24$$

(b)$\sigma_{c=50}(Y) \bowtie Z$.

$$\frac{T(Y) \times T(Z)}{V(Y,c) \times \mathrm{Max}(V(Y,d), V(Z,d))} = \frac{600 \times 400}{25 \times \mathrm{Max}(50,80)} = 125$$

5) Assume A=10,B=20 (here we imagine A and B are blocks that can hold 1 integer) are stored in a DB. Suppose a transaction does the following sequence of operations I(A), I(B), R(A,a), R(B,b), a:= a+b, b:=b+ 2*b, W(A,a), W(B,b), O(A), O(B). Show the undo log records needed for this transaction.

Transaction Op	Value a	Value b	Mem Value A	Disk Value A	Mem Value B	Disk Value B	Log Records
							<START T>
I(A)			10	10		20	
I(B)			10	10	20	20	
R(A,a)	10		10	10	20	20	
R(B,b)	10	20	10	10	20	20	
a:= a+b	30	20	10	10	20	20	
b:= a+2*b	30	70	10	10	20	20	
W(A,a)	30	70	30	10	20	20	<T, A, 10>
W(b,b)	30	70	30	10	70	20	<T, B, 20>
O(A)	30	70	30	30	70	20	
O(B)	30	70	30	30	70	70	
							<COMMIT T>
FLUSH LOG							